

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **07056945 A**

(43) Date of publication of application: **03.03.95**

(51) Int. Cl.

G06F 17/30

G06F 12/00

(21) Application number: **05204351**

(22) Date of filing: **18.08.93**

(71) Applicant: **TOPPAN PRINTING CO LTD**

(72) Inventor:
UEKI HIROMI
MAKINOUCHI KOUJI
HIRASAWA MICHIIKO
NARA MASAHITO
HAMAYA GUNJI

(54) **WHOLE SENSITIVE DATA BASE SYSTEM**

COPYRIGHT: (C)1995,JPO

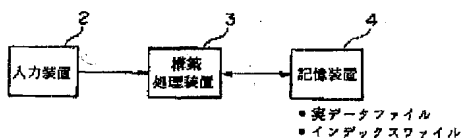
(57) Abstract:

PURPOSE: To provide a whole sentence data base system where the size of an index file is set to be small and sufficient retrieval speed can be obtained.

CONSTITUTION: This system is provided with an input device 2 inputting document data, a construction processor 3 generating a Japanese sentence lowest layer table J2-1 and a Japanese sentence highest layer table J2-2, which have peculiar pseudo words whose character string length is more than '2', in terms of hierarchy based on document data supplied from the input device 2, a storage device 4 storing document data and the respective tables J2-1 and J2-2 as a real data file and an index file, an input device 6 inputting a retrieval character string and a retrieval processor 7 extracting the peculiar pseudo word which agrees with a retrieval pseudo word whose character string length constituting the retrieval character string becomes '2' from the index file and outputting document data corresponding to the retrieval character string to a display 9.

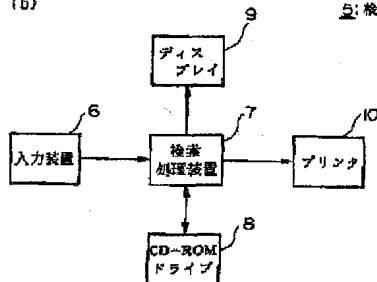
(a)

1:構築システム



(b)

5:検索システム



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-56945

(43) 公開日 平成7年(1995)3月3日

(51) Int.Cl. ⁸	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/30 12/00	5 2 0 A	8944-5B 9194-5L 9194-5L	G 0 6 F 15/ 40 15/ 413	3 7 0 A 3 1 0 A

審査請求 未請求 請求項の数 5 O L (全 15 頁)

(21) 出願番号 特願平5-204351

(22) 出願日 平成5年(1993)8月18日

(71) 出願人 000003193

凸版印刷株式会社

東京都台東区台東1丁目5番1号

(72) 発明者 植木 広実

東京都台東区台東一丁目5番1号 凸版印刷株式会社内

(72) 発明者 牧之内 浩二

東京都台東区台東一丁目5番1号 凸版印刷株式会社内

(72) 発明者 平澤 道彦

東京都台東区台東一丁目5番1号 凸版印刷株式会社内

(74) 代理人 弁理士 志賀 正武 (外2名)

最終頁に続く

(54) 【発明の名称】 全文データベースシステム

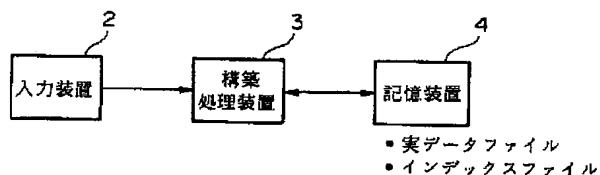
(57) 【要約】

【目的】 インデックスファイルのサイズを小とし、十分な検索速度を得ることができる全文データベースシステムを提供する。

【構成】 文書データを入力する入力装置2と、入力装置2から供給される文書データに基づいて、文字列長が2以上である固有疑似単語を有する和文最下層テーブルJ2-1および和文最上層テーブルJ2-2等を階層的に作成する構築処理装置3と、文書データおよび各テーブルJ2-1、J2-2等を実データファイルおよびインデックスファイルとして記憶する記憶装置4と、検索文字列を入力する入力装置6と、検索文字列を構成する文字列長が2となる検索用疑似単語に一致する固有疑似単語を上記インデックスファイルから抽出し、検索文字列に対応する文書データをディスプレイ9へ出力する検索処理装置7とから構成される。

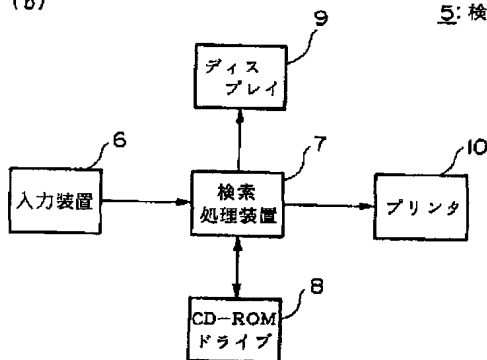
(a)

1: 構築システム



(b)

5: 検索システム



【特許請求の範囲】

【請求項1】 階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、

前記文書データ中の各文字と後続する文字とで構成される合計k文字（kは2以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、

固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、

前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段とを具備することを特徴とする全文データベースシステム。

【請求項2】 前記インデックステーブル群は、合計k文字未満の疑似単語に対応したテーブルをも有することを特徴とする請求項1に記載の全文データベースシステム。

【請求項3】 検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、連続するアドレスが付与された前記文書データ中の各文字、および当該各文字に後続する文字から構成される合計k文字（kは2以上）の各疑似単語の先頭文字列に付与された各アドレスを、対応する疑似単語の文字コード順に記憶するアドレステーブルと、

階層構造を有するテーブル群であって、固有の文字コードを有する疑似単語を固有疑似単語として文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスが対応付けられる最下層テーブル、この最下層テーブルの上層に構築される複数の上層テーブルからなり、各上層テーブルは1段下層のテーブルを略均等に分割する位置に記憶された固有疑似単語を文字コード順に記憶する階層構造をなし、最上層テーブルに記憶される固有疑似単語の数が予め設定された数以下となるよう構成されるインデックステーブル群と、検索文字列を入力する入力手段と、

該入力手段から供給される前記検索文字列をk文字単位に分割し、複数の検索用疑似単語を生成する分割手段と、

前記各検索用疑似単語と文字コードが同一である固有疑似単語を前記インデックステーブル群から抽出する抽出

処理を行い、抽出された各固有疑似単語に対応する各アドレスの差から前記文書データ中で連続して存在する固有疑似単語の組を特定するとともに、前記組に対応するアドレス群に応じた文書データ中の文字列を出力する検索手段とを具備し、

前記検索手段は、前記検索用疑似単語と文字コードが一致する固有疑似単語が検索対象となるテーブル中に存在しない場合には、前記検索用疑似単語より小なる文字コードの固有疑似単語群から最も前記検索用疑似単語に近い文字コードの最小疑似単語と、この最小疑似単語の直後に記憶される最大疑似単語とを抽出するとともに、1段下層のテーブルを検索対象とし、当該テーブルにおいて、前記最小疑似単語に一致する文字コードの固有疑似単語と、前記最大疑似単語に一致する文字コードの固有疑似単語とに挟まれる範囲に対して前記抽出処理を施すことを特徴とする全文データベースシステム。

【請求項4】 階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、

前記文書データ中の各文字と後続する文字とで構成される合計k文字（kは2以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、

固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、

前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段と、

前記検索文字列を入力する入力手段と、

該入力手段から供給される前記検索文字列をk文字単位に分割し、複数の検索用疑似単語を生成する分割手段と、

前記インデックステーブル群から前記検索用疑似単語と文字コードが同一である固有疑似単語を抽出する抽出処理を行い、抽出された各固有疑似単語に対応する各アドレスの差から前記文書データ中で連続して存在する固有疑似単語の組を特定するとともに、前記組に対応するアドレス群に応じた文書データ中の文字列を出力する検索手段とを具備し、

前記検索手段は、前記検索用疑似単語と文字コードが一致する固有疑似単語が検索対象となるテーブル中に存在

10

20

30

40

50

しない場合には、前記検索用疑似単語より小なる文字コードの固有疑似単語群から最も前記検索用疑似単語に近い文字コードの最小疑似単語と、この最小疑似単語の直後に記憶される最大疑似単語とを抽出するとともに、1段下層のテーブルを検索対象とし、当該テーブルにおいて、前記最小疑似単語に一致する文字コードの固有疑似単語と、前記最大疑似単語に一致する文字コードの固有疑似単語とに挟まれる範囲に対して前記抽出処理を施すことを特徴とする全文データベースシステム。

【請求項5】 前記インデックステーブル群は、合計k文字未満の疑似単語に対応したテーブルをも有し、前記検索手段は前記検索文字列の長さに応じて前記検索対象とするテーブルを変更することを特徴とする請求項3または4に記載の全文データベースシステム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は電子出版等に用いて好適な全文データベースシステムに関する。

【0002】

【従来の技術】文書データから所定の文字列を検索する手法には、大別して、キーワード検索と全文検索とがある。キーワード検索では、キーワード登録者が、文書データを例えば、文章毎に細分化し、各文書データに適合するキーワードを対応させる。この対応関係はインデックスファイルに記憶され、当該ファイルにおいて、各キーワードは例えば、文字コード順に木構造をなすように記憶される。

【0003】このような構成において、検索操作者が抽出したい文書データに対応するキーワードを入力することにより検索処理が為される。具体的には、入力キーワードに一致するキーワードがインデックスファイルから抽出され、このキーワードに対応する文書データが出力される。

【0004】一方、全文検索では、文書データはシーケンシャルファイル（全文データベース）として記憶される。この文書データに対する検索処理は、検索操作者が、抽出したい文書データに含まれる文字列を入力することにより為される。具体的には、入力された検索文字列と文書データに含まれる全ての文字列とが比較され、一致する文字列が含まれる文書データが出力される。なお、検索文字列の文字数や適用するパターンマッチングの手法によっては、文書データに含まれる全ての文字列を検索文字列と比較する必要はない。各種パターンマッチングの手法については、公知であるので、その説明を省略する。

【0005】

【発明が解決しようとする課題】一般に、データベース検索を必要とする電子出版の分野において、取り扱われる文書データの容量は、極めて大きなものとなる。例えば、CD-ROMは数百MB（メガバイト）の記憶容量

を有しており、通常、数十～数百MBの文書データが記憶される。このように、極めて大容量の文書データに対して、前述した全文検索を適用すると、公知のいかなるパターンマッチングの手法を用いても、パーソナルコンピュータ程度の処理能力では、十分な応答時間を達成することができないという欠点がある。

【0006】また、前述したキーワード検索を適用しようとしても、CD-ROMに記憶される文書データは、データベース化を前提として作成されていない為に、文章毎に細分化する作業や、適合するキーワードを作成する作業が極めて困難になるという問題がある。さらに、キーワード登録者と検索操作者は別人であるために、検索操作者が、作成されたキーワードを有効に活用することが困難であるという欠点もある。

【0007】また、使用者が必要とする情報は、「経済」、「製法」、「演算」等の数文字の文字列により特定できる場合が多く、これらの文字列の中には、キーワードとして登録しにくいものが含まれることがある。すなわち、登録されたキーワード以外の文字列による検索が必要となる場合がある。こうした場合に対応できるように、例えば、2文字からなる文字列を全てキーワードとすることも考えられる。

【0008】しかしながら、日本語の場合、約7千種の文字が存在するために、2文字の順列は約4千9百万という膨大な数となる。この膨大な数のキーワードから特定のキーワードを抽出する処理は極めて高負荷となり、木構造を用いても、パーソナルコンピュータ程度の処理能力では実用にならない。また、検索の為のインデックスファイルも巨大（例えば、数百MB）なものとなり、文書データの為の記憶領域が狭くなってしまう。

【0009】本発明は、このような背景に鑑みて為されたもので、インデックステーブル群のサイズを小とし、十分な検索速度を得ることができる全文データベースシステムを提供することを目的とする。

【0010】

【課題を解決するための手段】本発明による全文データベースシステムは、階層構造を有するインデックステーブル群を用いて、検索文字列に一致する文字列を文書データから抽出する全文データベースシステムであって、前記文書データ中の各文字に連続するアドレスを付与するアドレス付与手段と、前記文書データ中の各文字と後続する文字とで構成される合計k文字（kは2以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶するアドレステーブル作成手段と、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける最下層構築手段と、前記インデックステーブル群中の最上層

テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する階層化手段とを具備することを特徴としている。

【0011】

【作用】上記構成によれば、アドレス付与手段が、文書データ中の各文字に連続するアドレスを付与し、アドレステーブル作成手段が、前記文書データ中の各文字と後続する文字とで構成される合計k文字(kは2以上)の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶する。そして、最下層構築手段が、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける。さらに、階層化手段が、前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する。インデックステーブル群は、このような階層構造を有する為に、検索時において、十分な検索速度が得られる。また、インデックステーブル群中の各テーブルに記憶される疑似単語は、固有の文字コードを有する固有疑似単語であるために、インデックステーブル群のサイズが小となる。

【0012】

【実施例】以下、図面を参照して、本発明の一実施例について説明する。

(1) 構築システム1の構成

図1は本発明の一実施例による全文データベースシステムの概略構成を示す図であり、図1(a)は全文データベースを構築する構築システム1の概略構成を示すブロック図である。この構築システム1は、データベースの提供者(もしくは編集者)に使用されることが想定される。電子出版においては、その提供者が当該システム1*

$$CC = C1 \times m + C2 \quad \dots (1)$$

文書データ中に現れる文字種は約7千種であるので、この文字種以上の数「m(例えば、8千)」をC1に乘ずることにより、CCは疑似単語固有の文字コードとなる。

【0017】また、各レコードR2は、当該レコードR2に格納される固有疑似単語の文書データ中での存在数を表す「サイズ」を有し、文字アドレステーブルCAT中の所定のレコードR1に対応付けられている。ここで、レコードR2が対応付けられるレコードR1は、当該レコードR2が有する固有疑似単語の先頭文字の位置

*により構築された全文データベースを、CD-ROM等に記憶させる。

【0013】図1(a)において、2はキーボード等の入力装置であり、オペレータにより入力される文書データを構築処理装置3(後述する)へ供給する。構築処理装置3は、入力装置2から供給される文書データに所定の処理を施して、記憶装置4(後述する)へ供給するとともに、当該文書データに対応するインデックスファイルを作成し、全文データベースを構築する。この全文データベース構築処理の内容については、後に詳述する。

【0014】記憶装置4は、例えば、数百MBの容量を有するハードディスクからなり、構築処理装置3から供給される文書データを実データファイルとして記憶するとともに、構築処理装置3にて作成されるインデックスファイルを記憶する。ここで、実データファイル中の文書データの一例を図4に示す。この図に示す文書データには、キーワード“東京都”および“京都”を識別する為の位置マーク‘@’が付加されている。

【0015】また、インデックスファイルに含まれる各種テーブルの一例を図12に示す。図12は図4に示す文書データの和文範囲JAに対するインデックステーブル(以後、和文テーブルと称す)の構成例を示す図であり、この図において、J2-1は和文最下層テーブル、J2-2は和文最下層テーブルJ2-1の上層テーブルとなる和文最上層テーブル、J1は1文字テーブルであり、和文最下層テーブルJ2-1の上層テーブルとなる。また、CATはレコードR1を複数有する文字アドレステーブルである。

【0016】和文最下層テーブルJ2-1はレコードR2を複数有し、各レコードR2には、文書データの和文範囲JA(図4参照)中に存在する2文字の文字列(以後、疑似単語と称す)が格納されている。これらの疑似単語は、ユニーク(同一綴りのものが無い)であり、文字コードCC順にソートされている(以後、ユニークな疑似単語を固有疑似単語と称す)。ここで、疑似単語の文字コードCCは、当該疑似単語の先頭文字の文字コードをC1、末尾文字の文字コードをC2とすると、例えば、以下に示す計算式(1)により算出される。

を表す「文字アドレス」を有する。

【0018】なお、和文最下層テーブルJ2-1において、サイズが複数であるレコードR2には、複数のレコードR1が対応付けられる。具体的には、サイズが複数であるレコードR2に、当該レコードR2が有する固有疑似単語に対応する複数の文字アドレスのうち、最小の文字アドレスを有するレコードR1が対応付けられ、このレコードR1に続いて、最小でない文字アドレスを有するレコードR1が昇順に整理される。

【0019】また、和文最上層テーブルJ2-2はレコ

ードR3を複数有し、各レコードR3には、和文最下層テーブルJ2-1中の固有疑似単語が格納される。なお、このレコードR3の数は、レコードR2に較べて極めて少ない。また、各レコードR3は、固有疑似単語の文字コード順にソートされており、レコードR3が有する固有疑似単語と同一の固有疑似単語を有するレコードR2に対応付けられる。この際、和文最下層テーブルJ2-1において、各レコードR3に対応付けられる各レコードR2間の距離、すなわち、当該レコードR2間に存在するレコードR2の数が、予め設定された数（例えば、99）となるように、各レコードR3に格納される固有疑似単語が設定される。

【0020】1文字テーブルJ1はレコードR4を複数有し、各レコードR4は、文書データの和文範囲JAに出現する全種類の文字を格納する。これらの文字はユニークであり、各レコードR4は文字コード順にソートされている。また、各レコードR4は、和文最下層テーブルJ2-1中の所定のレコードR2に対応付けられており、対応付けられるレコードR4およびレコードR2が有する文字および固有疑似単語の先頭文字は一致する。ここで、1つのレコードR4が対応付けられるべきレコードR2が複数である場合には、当該レコードR4は最も先頭にあるレコードR2に対応付けられる。この際、レコードR4が有するサイズは、対応付けるべき固有疑似単語の数となる。

【0021】次に、文書データ（図4参照）の欧文範囲EAに対応するインデックステーブル（以後、欧文テーブルと称す）の一例を図7および図9に示す。欧文テーブルは、図7に示すようなワードアドレステーブルWAT、ワードリストWLと、図9に示すような仮想アドレステーブルVAT、欧文基本テーブルEBTとからなる。図7に示すように、ワードアドレステーブルWATはレコードR5を複数有し、各レコードR5には、欧文範囲EAに存在する全ての「ワード（単語）」に対応するワード単位のアドレス（以後、ワードアドレスと称す）が格納される。

【0022】また、ワードリストWLは、レコードR6を複数有し、各レコードR6は、欧文範囲EAに存在するユニークなワード（以後、固有ワードと称す）を1つずつ有し、各固有ワードの文字コード順にソートされている。また、各レコードR6には、固有の「ユニーク符号」が割り当てられ、各固有ワードを構成する各文字には、固有ワード内の先頭からの位置を示す「ワード内アドレス」が付与される。

【0023】また、各レコードR6は、自身が有する固有ワードが文書データ中で出現する頻度を表す「サイズ」を有し、ワードアドレステーブルWAT中の所定のレコードR5に対応付けられる。なお、サイズが複数であるレコードR6には、複数のレコードR5が対応付けられる。具体的には、サイズが複数であるレコードR6

に、当該レコードR6が有する固有ワードに対応する複数のワードアドレスのうち、最小のワードアドレスを有するレコードR5が対応付けられ、このレコードR5に続いて、最小でない文字アドレスを有するレコードR5が昇順に整列される。

【0024】また、図9に示す欧文基本テーブルEBTは、レコードR7を複数有し、各レコードR7には、ワードリストWL（図7参照）中の各固有ワードに基づいて生成される2文字単位の固有疑似単語が1つずつ格納される。各レコードR7は、固有疑似単語の文字コード順にソートされている。各レコードR7は、格納された固有疑似単語のワードリストWL中での出現頻度を表す「サイズ」を有し、仮想アドレステーブルVAT中の所定のレコードR8に対応付けられる。

【0025】仮想アドレステーブルVATはレコードR8を複数有し、各レコードR8には、ユニーク符号とワード内アドレスとの組である「仮想アドレス」が格納される。この仮想アドレスは、欧文基本テーブルEBT中の固有疑似単語に対応するものである。ここで、レコードR7のサイズが複数である場合には、当該レコードR7に複数のレコードR8が対応付けられる。この対応付けの具体的内容は、図12に示す和文最下層テーブルJ2-1のレコードR2と、文字アドレステーブルCATのレコードR1との対応付けと同様であるので、その説明を省略する。

【0026】(2) 検索システム5の構成

一方、図1(b)は構築システム1により構築された全文データベースに対して、各種の検索処理を行う検索システム5の構成を示す図である。この検索システム5は、データベースのユーザーによって用いられることが想定される。電子出版においては、そのユーザーが当該システム5を用いて、CD-ROM等に記憶された全文データベースに対して各種の検索処理を行う。

【0027】図1(b)において、6はキーボード等の入力装置であり、ユーザー（使用者）により入力される指示に応じた指示データを検索処理装置7（後述する）へ供給する。検索処理装置7は、一般的なパーソナルコンピュータであり、図示せぬCPU、ROM、RAMおよび各種I/Oインタフェースを有する。この検索処理装置7はRAMに記憶される検索プログラムを実行し、入力装置6から供給される指示データに応じた検索処理を行う。この検索処理の内容は後に詳述する。

【0028】8はCD-ROMドライブであり、検索処理装置7に制御され、挿入されるCD-ROMに記憶された情報を読み取る。9は検索処理装置7から供給される表示データに応じて、検索メニューや検索結果等を表示するディスプレイ、10はプリンタであり、検索処理装置7から供給される出力データに応じて、検索結果を出力する。

【0029】(3) 全文データベース構築処理

次に、構築処理装置3(図1(a)参照)がRAMに記憶されたプログラムを実行して行う全文データベース構築処理について、以下に説明する。ここでは、和文と欧文が混在した文書データ(図3参照)をデータベース化する場合について説明する。

【0030】まず、全文データベース構築に先だって、図1(a)に示す構築システム1において、入力装置2から文書データが入力される。この文書データは、構築処理装置3を介して記憶装置4に供給され、ここで記憶される。以下に説明する各処理は、記憶装置4に記憶された文書データに対して為される。文書データの入力処理が終了し、入力装置2から所定の指示データが入力されると、構築処理装置3は、図2のフローチャートに表されるプログラムを実行する。

【0031】まず、ステップSA1では、文書データ(図3参照)に位置マークを付加する。この位置マークとは、任意のキーワードの前後に挿入される特定の文字であり、例えば、' @ ' のように、文書データ中に存在しない記号を用いる。図4は、上述したマーク付加処理が行われた後の文書データを示す図であり、この図において、" 東京都 " および " 京都 " という文字列(キーワード)の前後には位置マーク ' @ ' が挿入されている。

【0032】上述したマーク付加処理を行わない場合、" 京都 " という地名を全文検索すると、" 京都 " はもちろん、" 東京都 " まで抽出してしまう。マーク付加処理は、このような無意味な抽出を避ける為に行われる処理であり、全文検索において、検索文字列として " @ 京都 @ " という文字列を入力すると、" @ 京都 @ " のみが抽出され、" @ 東京都 @ " は抽出されないという結果を得ることができる。

【0033】次に、ステップSA2(図2参照)では、文書データにアドレスが付与される。図5に示すように、アドレスには文字アドレスとワードアドレスがあり、文字アドレスは文書データ全体に付与され、ワードアドレスはアルファベットや数字等が連続する欧文範囲EAに付与される。例えば、図5の和文範囲JAにおいて、先頭文字 ' 多 ' の文字アドレスは「1」、それに続く文字 ' 角 ' の文字アドレスは「2」となり、欧文範囲EAにおいて、先頭文字 ' w ' の文字アドレスは「317」となる。また、欧文範囲EAにおいて、先頭のワード " world " のワードアドレスは「1」、それに続くワード " wide " のワードアドレスは「2」となる。

【0034】後述する各処理において、欧文範囲EAにはワード単位の処理が行われるため、欧文範囲EAにおいて、文字アドレスを記憶する必要はない。しかしながら、和文範囲JAとの位置関係を把握するために、欧文範囲EAの最初および最後の文字アドレスを記憶装置4(図1(a)参照)に記憶する。これらの文字アドレス間の文書データは、後述する各処理において、欧文範囲

EAとみなされ、ワード単位の処理を施される。

【0035】次に、ステップSA3(図2参照)では、欧文範囲EAのワードリストWLを作成する。まず、図5の文書データの欧文範囲EAに出現する全てのワードを抽出し、図6に示すように、欧文対照テーブルCTを作成する。次に、欧文対照テーブルCT中の各レコードR11を、各ワードの文字コード順およびワードアドレス順にソートする。すると、同一のワード(例えば、図6中のワード " world " 参照)を有する複数のレコードR11が隣接する。

【0036】次に、ソートされた各レコードR11からユニークな固有ワードを抽出し、ワードリストWL(図7参照)の各レコードR6に格納するとともに、ワードアドレスを有するレコードR5を複数作成し、ワードアドレステーブルWATを作成する。ワードリストWLの各レコードR6に設けられるポインタは、ワードアドレステーブルWAT中のレコードR5を指し示す。

【0037】この際、ポインタにより対応付けられるレコードR6の固有ワードとレコードR5のワードアドレスとは、欧文対照テーブルCTにおいて同一レコードR11内に格納されていたもの同士となる。また、欧文対照テーブルCTにおいて、同一のワードを有するレコードR11が複数存在していた場合、そのワードに一致する固有ワードを有するワードリストWL中のレコードR6には、当該ワードの欧文対照テーブルCT内での出現数がサイズとして格納される。例えば、ワード " help " は、欧文対照テーブルCT(図6参照)中に2つ出現するので、ワードリストWLの固有ワード " help " を有するレコードR6のサイズは「2」となる。

【0038】さらに、ワードリストWLの各レコードR6には、固有のユニーク符号「A」、「B」、「C」、・・・が付与され、各レコードR6の固有ワードを構成する文字には、ワード内アドレスが付与される。例えば、固有ワード " can " に付与されるユニーク符号は「A」であり、固有ワードを構成する文字 ' c ' に対するワード内アドレスは「1」である。こうして作成されたワードリストWLは、記憶装置4(図1(a)参照)に記憶される。

【0039】次に、ステップSA4(図2参照)では、文書データまたはワードリストWLから疑似単語を抽出する。これに続くステップSA5では、抽出された疑似単語を用いて、和文最下層テーブルJ2-1および欧文基本テーブルEBTを作成する。これらの抽出処理および作成処理は、文書データの形式により異なる為、以下、欧文範囲EAと和文範囲JAとに分けて説明する。

【0040】A: 欧文範囲EAに対する処理

欧文範囲EAにおいて、まず、ワードリストWL(図7参照)から、文字列長が「2」である疑似単語を抽出する。この抽出処理は、各固有ワードの先頭から末尾にかけて行われ、例えば、固有ワード " can " からは、

10

20

30

40

50

ca", "an" という疑似単語が抽出される。こうして抽出された複数の疑似単語は、図8に示すような構成の疑似単語テーブルPWTの各レコードR9に格納される。

【0041】疑似単語テーブルPWTにおいて、疑似単語が格納されたレコードR9は、当該レコードR9が有する疑似単語の抽出元の固有ワードに付与されたユニーク符号と、その疑似単語の先頭文字のワード内アドレスとから構成される「仮想アドレス」を有する。例えば、疑似単語"an"を有するレコードR9は、抽出元の固有ワード"can"に付与されたユニーク符号「A」と、疑似単語"an"の先頭文字'a'のワード内アドレス「2」とから構成される仮想アドレス「A-2」を有する。

【0042】そして、ワードリストWL作成時と同様に、疑似単語テーブルPWT内の各レコードR9を各疑似単語の文字コード順にソートし、ユニークな固有疑似単語を抽出する。ここで、抽出された固有疑似単語は、図9に示す欧文基本テーブルEBTに格納される。また、疑似単語テーブルPWT内の仮想アドレスのみで構成される仮想アドレステーブルVATを作成する。図9に示すように、欧文基本テーブルEBTの各レコードR7に設けられるポインタは、仮想アドレステーブルVAT中の対応するレコードR8を指し示す。

【0043】また、ワードリストWL作成時と同様に、疑似単語テーブルPWTにおいて、同一の疑似単語が複数存在していた場合、その疑似単語と同一の固有疑似単語を有するレコードR7には、その疑似単語の出現数が「サイズ」として格納される。こうして、図9に示す欧文基本テーブルEBTが作成される。

【0044】B: 和文範囲JAに対する処理

和文範囲JAにおいては、まず、文書データ(図5参照)から、文字列長が「2」である疑似単語を抽出する。この抽出処理は、和文範囲JAの先頭から末尾にかけて行われ、例えば、図5の文書データからは、順に"多角", "角経" という疑似単語が抽出される。抽出された疑似単語は、図10に示すような和文疑似単語テーブルJPTの各レコードR10に格納される。

【0045】和文疑似単語テーブルJPTにおいて、疑似単語が格納されたレコードR10は、当該疑似単語の先頭文字の「文字アドレス」を有する。例えば、疑似単語"多角"を有するレコードR10は、疑似単語の先頭文字'多'の文字アドレス「1」を有する。そして、ワードリストWLや疑似単語テーブルPWT作成時と同様に、各レコードR10を疑似単語の文字コード順にソートし、ユニークな固有疑似単語を抽出する。

【0046】ここで、抽出された固有疑似単語は、図11に示す和文最下層テーブルJ2-1に格納される。また、和文疑似単語テーブルJPT内の文字アドレスのみで構成される文字アドレステーブルCATを作成する。

図11に示すように、和文最下層テーブルJ2-1の各レコードR2に設けられるポインタは、文字アドレステーブルCAT中の対応するレコードR1を指し示す。

【0047】また、和文疑似単語テーブルJPT(図10参照)において、同一の疑似単語が複数存在していた場合、その疑似単語に一致する固有疑似単語を有するレコードR2には、当該疑似単語の文書データ中での出現数が「サイズ」として格納される。こうして、図11に示す和文最下層テーブルJ2-1が作成される。上述したように、欧文基本テーブルEBTおよび和文最下層テーブルJ2-1が作成されると、処理はステップSA6へ進む。

【0048】ステップSA6において、最上層の和文テーブル(最初は和文最下層テーブルJ2-1)中のレコード数が所定数n(例えば、n=500)以上であるか否かが判断される。この判断が「Yes」であれば、処理はステップSA7へ進み、「No」であれば、処理はステップSA8へ進む。

【0049】上記ステップSA6での判断により、必要に応じて、和文テーブルが階層化されるのだが、ここで、当該階層化を行う理由を説明する。前述したように、和文に用いられる文字種は約7千種と多く、文書データから抽出される固有疑似単語の種類、すなわち、和文最下層テーブルJ2-1のレコードR2の数は極めて大となる。後述する検索処理は、検索文字列に対応する固有疑似単語を抽出する処理を繰り返して行われるため、検索対象となるテーブル(例えば、和文最下層テーブルJ2-1)のレコードR2の数が多いと、所定の時間内に検索処理を終了することができない。

【0050】ここで、和文最下層テーブルJ2-1のレコードR2の数が多き場合には、図12に示すように、上層のテーブル(和文最上層テーブルJ2-2)を作成し、上層のテーブルで検索文字列に対応する固有疑似単語を抽出できなかった場合には、下層テーブル(和文最下層テーブルJ2-1)の所定の範囲で、固有疑似単語を抽出するようにする。すると、比較すべき固有疑似単語の数が減少し、所定の応答時間で検索処理を行うことができる。

【0051】なお、欧文基本テーブルEBTに関して上記階層化を行わないのは、当該テーブルEBT中の各固有疑似単語は、文字種の少ないアルファベットや数字等の組み合わせであるために、そのレコード数は、和文最下層テーブルJ2-1のレコード数に比べて極めて少なく(同一綴りの疑似単語が多い)、欧文基本テーブルEBTのみでも十分な応答時間を得ることができるからである。

【0052】ステップSA7は、ステップSA6での判断が「Yes」となった場合の処理であり、ここでは、既に作成された和文テーブルに対して上層の和文テーブルを作成する。例えば、図10の和文最下層テーブルJ

2-1に対して、図11に示すように、和文最上層テーブルJ2-2を作成する。この和文最上層テーブルJ2-2の各レコードR3には、“@東”や“経常”等の和文最下層テーブルJ2-1から抽出された固有疑似単語が格納される。また、各レコードR3は「ポインタ」を有し、和文最下層テーブルJ2-1中の同一固有疑似単語を有するレコードR2に対応付けられる。

【0053】なお、各レコードR3が有する固有疑似単語は、隣接するレコードR3に対応する各レコードR2間の距離（2つのレコードR2に含まれるレコードR2の数）が、例えば、99となるように抽出される。この距離は和文最下層テーブルJ2-1のレコード数に応じて設定される。そして、処理はステップSA6に戻る。こうして、最上層の和文テーブルのレコード数が所定数n未満となるまで、上述した階層化処理が行われる。図12に示す例では、和文最上層テーブルJ2-2のレコード数は所定数n（例えば、n=500）未満となるので、和文テーブルの階層は2段となる。

【0054】ステップSA8は、最上層の和文テーブルのデータ数が所定数n未満となり、ステップSA6での判断が「No」となる場合の処理であり、和文最下層テーブルJ2-1に対する1文字テーブルJ1が作成される。和文には漢字が用いられるため、「山」や「川」等の1文字の検索文字列による検索が行われる場合がある。こうした1文字検索をも所定の応答速度で実現する為に、1文字テーブルJ1が作成される。

【0055】1文字テーブルJ1の作成過程を以下に説明する。まず、図12に示す和文最下層テーブルJ2-1から各固有疑似単語の先頭文字をレコード順に抽出する。各レコードR2は、既に文字コード順にソートされている為、同一の文字が連続して抽出される。次に、抽出された文字群からユニークな文字を抽出し、抽出元の文字群に含まれる同一文字の数（サイズ）とともに、1文字テーブルJ1の各レコードR4に格納する。

【0056】また、各レコードR4はポインタを有し、和文最下層テーブルJ2-1内の抽出元レコードR2に対応付けられる。こうして、1文字テーブルJ1が作成される。そして、例えば、図7のワードアドレステーブルWAT、ワードリストWLと、図9の仮想アドレステーブルVAT、欧文基本テーブルEBTと、図12の文字アドレステーブルCAT、和文最下層テーブルJ2-1、和文最上層テーブルJ2-2、1文字テーブルJ1と、欧文範囲EAの最初および最後の文字アドレスとが、インデックスファイルとして、記憶装置4に記憶される。また、図5に示すような文書データが実データファイルとして記憶装置4に記憶され、全文データベースが構築される。

【0057】(4) 全文検索処理

次に、上述した過程を経て構築された全文データベースに対して、検索処理装置7（図1（b）参照）が行う全

文検索処理について、図面を参照して説明する。図13、図14は検索処理装置7のRAMに予め記憶される全文検索プログラムのフローチャートである。まず、検索システム5において、CD-ROMドライブ8に、全文データベースが記憶されたCD-ROMが挿入され、入力装置6から所定の指示データが供給されると、検索処理装置7はステップSB1を実行する。

【0058】ステップSB1では、所定の表示データをディスプレイ9へ供給し、例えば、図15に示す検索メニューを表示させる。ユーザーは、表示された検索メニューに応じて、入力装置6を操作し、後述する検索モード、順位モードおよび指定距離等を設定する。検索モードには1つの検索文字列を検索する通常検索モードの他に、複数の検索文字列を文脈上の関係を意識して検索する文脈意識モードがあり、ユーザーは入力装置6を操作して入力フィールド11へ所定の文字を入力し、どちらかのモードを選択する。

【0059】分脈意識モードを選択した場合、ユーザーは複数の検索文字列間の前後関係を意識するか否か（順位モード）を指定する必要がある。また、分脈意識モードでは、複数の検索文字列間の距離（先頭文字アドレスの差）の上限を指定する必要がある。したがって、ユーザーは入力フィールド12に順位モードを指定する文字を入力し、入力フィールド13に距離の上限を表す数値（指定距離）を入力する。

【0060】次に、ステップSB2では、ユーザーが入力装置6を操作し、図14の文字列入力フィールド14、あるいは文字列入力フィールド15へ検索文字列を入力する。そして、入力装置6から所定の指示データが供給されると、検索処理装置7は、検索メニュー上の各入力フィールド11～15に入力された各種のデータを読み取り、これらのデータをRAMに記憶する。そして、処理はステップSB3へ進む。

【0061】ステップSB3では、検索文字列の各文字に文字アドレスを付与し、検索文字列を2文字単位に分割して、複数の検索用疑似単語を抽出する。例えば、図17に示す“経営危機”という検索文字列からは“経営”と“危機”という検索用疑似単語が抽出される。検索文字列が2文字以下の長さであれば、上記抽出処理は行われない。

【0062】次に、ステップSB4では、各検索用疑似単語に一致する固有疑似単語を有するレコードを、記憶装置4に記憶されたインデックスファイルから検索する。この検索処理は各検索用疑似単語の文字コードSCと、インデックスファイル中の各固有疑似単語の文字コードVCとを比較することにより行われる。ここで、検索処理に使用されるテーブルは、欧文範囲EAでの検索では欧文基本テーブルEBT、和文範囲JAでの検索では最上層の和文テーブル（例えば、和文最上層テーブルJ2-2）あるいは1文字テーブルJ1となる。そし

て、ステップSB5では、上記検索処理が全ての検索用疑似単語に対して完了したか否かを判断する。この判断が「No」の場合はステップSB6へ、逆に「Yes」の場合はステップSB9へ処理が進む。

【0063】ステップSB6は、検索用疑似単語に対する検索処理が完了しなかった場合の処理であり、検索対象となっている和文テーブルが最下層のテーブル（例えば、和文最下層テーブルJ2-1）であるか否かを判断する。この判断が「No」の場合はステップSB7へ、逆に「Yes」の場合はステップSB18（図14参照）へ処理が進む。

【0064】ステップSB7は、検索対象となっている和文テーブルが、さらに下層のテーブルを有する場合の処理である。ここでは、検索対象となっている和文テーブル（例えば、和文最上層テーブルJ2-2）において、検索用疑似単語の文字コードSCより小さく、最も文字コードSCに近い文字コードVCの固有疑似単語を有するレコード（以後、近似レコードと称す）を抽出する。

【0065】そして、検索対象とする和文テーブルを1段下層のテーブル（例えば、和文最下層テーブルJ2-1）とし、このテーブルの特定範囲に対して、上層のテーブルに対する場合と同様な検索処理が施される。ここで、特定範囲とは、上層のテーブル中の近似レコードに対応付けられた下層テーブル中のレコード、および、このレコードに後続する99のレコードからなる。この検索処理が終了すると、処理はステップSB5へ戻る。ステップSB9は、ステップSB5での判断が「Yes」となる場合の処理であり、検索された各レコードが有する各種アドレスを抽出する。

【0066】検索されたレコードが和文テーブル（例えば、和文最上層テーブルJ2-2）に存在する場合は、当該レコードに対応付けられた最下層の和文テーブル（例えば、和文最下層テーブルJ2-1）中のレコードを抽出し、当該レコードに対応付けられる文字アドレスを抽出する。ここで抽出された最下層の和文テーブル中のレコードが有するサイズが複数である場合は、上記文字アドレスおよび後続する文字アドレス群から、順に、サイズの数だけ文字アドレスを抽出する。

【0067】また、検索文字列が1文字である場合には、1文字テーブルJ1中の抽出されたレコードに対応付けられる最下層の和文テーブル中のレコードを抽出する。この際、1文字テーブルJ1中の検索されたレコードのサイズが複数であれば、当該レコードに対応付けられた和文最下層テーブルJ2-1のレコードおよびこのレコードに後続するレコード群から、順に、サイズの数だけレコードを抽出する。こうして抽出された和文最下層テーブルJ2-1中の各レコードに対応する文字アドレスを抽出し、昇順にソートする。

【0068】次に、ステップSB10では、各検索用疑

似単語に対応して抽出された文字アドレス群のうち、各検索用疑似単語間の距離に相当する差を有する文字アドレスの組を抽出し、抽出された組の先頭アドレスを抽出する。例えば、図17に示すように、検索文字列が“経営危機”であれば、検索用疑似単語“経営”および“危機”間の距離は2である。

【0069】したがって、検索用疑似単語“経営”に対応して抽出された文字アドレスと、検索用疑似単語“危機”に対応して抽出された文字アドレスとの差が2となる組を抽出する。この際、各検索用疑似単語に対応する文字アドレス群はソートされている為に、各々のアドレス群から小さい順に文字アドレスを抽出し、これらを比較することにより、差が2となる文字アドレスの組が容易に抽出される。そして、抽出された文字アドレスの組の先頭文字アドレス（“経営危機”の場合は“経”に対応する文字アドレス）が抽出される。

【0070】また、検索文字列がアルファベットであり、例えば、欧文基本テーブルEBTから1つあるいは複数のレコードR7が抽出された場合には、まず、当該レコードR7に対応付けられた仮想アドレステーブルVAT中のレコードR8を抽出する。そして、抽出されたレコードR8において、仮想アドレスのユニーク符号が同一のレコードR8について、ワード内アドレスの差が例えば、2となる仮想アドレスの組を抽出し、抽出された組の先頭文字アドレスを抽出する。

【0071】ここで、例えば、検索用文字列が“world”であれば、検索用疑似単語“wo”と“rl”との間隔は2、検索用疑似単語“rl”と“ld”との間隔は1である。したがって、検索用疑似単語“wo”に対応して抽出された仮想アドレス群と、検索用疑似単語“rl”に対応して抽出された仮想アドレス群とから、ユニーク符号が「A」であり、かつ、ワード内アドレスの差が2となる仮想アドレスの組を抽出し、こうして抽出された仮想アドレスと、検索用疑似単語“ld”に対応して抽出され、ユニーク符号が「A」である仮想アドレスとから、ワード内アドレスの差が1となる仮想アドレスの組を抽出する。

【0072】そして、抽出された組の仮想アドレスのうち、先頭の仮想アドレス中のユニーク符号からワードリストWL中のワードを特定する。特定されたワードには、ワードアドレスが対応付けられており、かつ、欧文範囲EAの先頭ワードには、和文範囲JAから連続する文字アドレス「317」も対応付けられているため、文書データ中における文字アドレスが得られる。もちろん、検索文字列が1文字である場合には、上述した連続性判断は行われない。

【0073】次に、ステップSB11では、分脈意識検索か否かが判断される。この判断が「Yes」であればステップSB12へ、「No」であればステップSB13へ処理が進む。ステップSB12は、分脈意識検索で

10

20

30

40

50

ある場合の処理である。分脈意識検索であれば、検索文字列は複数（ここでは、説明を簡略化するために2つとする。以後、各検索文字列を第1の検索文字列、第2の検索文字列と称す）であり、ここでは、第1および第2の検索文字列に対する検索処理が終了したか否かが判断される。この判断が「No」であれば、ステップSB3へ処理が戻り、未処理の検索文字列に対して上述した検索処理が施される。逆に、「Yes」であればステップSB13へ処理が進む。

【0074】ステップSB13では、第1の検索文字列に対応して抽出される先頭文字アドレス群と第2の検索文字列に対応して抽出される先頭文字アドレス群とが比較され、両者の差が指定距離以下となる文字アドレスの組（以後、範囲内アドレス組と称す）を抽出する。この際、2つの文字アドレスで規定される文書データ中に、キャリッジリターン等の区切り記号が存在する場合には、両者の差が指定距離以下であっても、範囲内アドレス組から除外される。

【0075】次に、ステップSB14では、範囲内アドレス組の数が1以上であるか否かが判断される。この判断が「Yes」であればステップSB15へ、「No」であればステップSB18へ処理が進む。ステップSB15では、順位指定があるか否かが判断される。この判断が「Yes」であればステップSB16へ、「No」であればステップSB17へ処理が進む。

【0076】ステップSB16は、順位指定があった場合の処理であり、範囲内アドレス組内の文字アドレスの順序が、指定された順序と一致する組（以後、順序一致アドレス組と称す）を抽出する。ここで抽出される組が0でない場合には、処理はステップSB17へ進む。逆に、当該組が存在しない場合には処理はステップSB18へ進む。

【0077】ステップSB17では、まず、各検索文字列に対応した先頭文字アドレス群に含まれる先頭文字アドレスの数に応じた表示データをディスプレイ9へ供給する。これにより、ディスプレイ9に表示されている検索メニューの出力フィールド16に、抽出されたデータ数が表示される。これを視認したユーザーが、入力装置6を操作し、所定の指示データを検索処理装置7へ供給すると、当該装置7は、先頭文字アドレス群中の文字アドレスを有する文書データに応じた表示データをディスプレイ9へ供給する。

【0078】こうして、図16に示すように、検索結果がディスプレイ9上に表示される。この際、文書データ中の検索文字列に一致する文字列は、例えば、反転表示され、他の文字列と区別される。また、分脈意識モードであった場合には、指定範囲外あるいは順位が一致しなかった先頭文字アドレスの文字列に下線が付される。ここで、ユーザーは、入力装置6を操作し、他の検索結果等をディスプレイ9上に表示させる。

【0079】また、ステップSB18は、ステップSB6、ステップSB14、あるいはステップSB16において、検索対象文字列に一致する文字列を検索できなかったと判断された場合の処理であり、ディスプレイ9へ所定の表示データを供給し、「指定された条件の検索文字列は文書中に存在しませんでした」等のメッセージを表示させる。

【0080】以上説明したように、本発明の一実施例によれば、文字列長が2の固有疑似単語を有する和文テーブルを階層的に構築する為に、和文テーブル自体のサイズを大きくすることなく、検索効率に優れた全文検索を行うことができる。また、1文字テーブルを設けた為に、文字列長が1の検索文字列に対する検索処理を迅速に行うことができる。さらに、ワードリストWLおよび欧文基本テーブルEBTを作成した為に、各固有疑似単語に対応するサイズを適度な大きさとすることができ、検索効率を向上させることができる。

【0081】また、欧文範囲EAにおいて、ワード単位よりも小さな疑似単語単位での検索が可能になるために、語尾変化したワードを一度に検索することができる。例えば、入力装置6を介して“econ”という検索文字列を入力すると、“economic”、“economy”というワードを抽出することができる。

【0082】なお、上述した一実施例においては、CD-ROMに全文データベースを記憶させる例を示したが、十分な記憶容量を有する記憶媒体であれば、CD-ROMでなくともよい。また、検索処理装置7はワークステーション等でも良く、パーソナルコンピュータである必要はない。さらに、1段下層のテーブルを分割する単位は99である必要はなく、固有疑似単語の数に応じて設定することが望ましい。あるいは、検索作業をそのレコードで終了させるストップレコードを挿入するようにしてもよい。

【0083】また、上述した一実施例においては、疑似単語の文字数を2文字として説明したが、2文字に限定されるものではなく、例えば、3文字、4文字、…というように複数文字であればよい。もちろん、疑似単語の文字数は、データベースの内容や検索処理の特徴等に応じて設定される。例えば、電子出版において、文書データが一般的な日本語である場合には、2文字程度に設定される。

【0084】さらに、上述した一実施例では、電子出版に適用する例を示した為に、構築システム1が全文データベースの提供者に使用され、検索システム5が全文データベースのユーザーに使用されるように、それぞれ個別のシステムとして構成されるが、両者を一体のシステムとして構成し、電子出版以外の分野で用いられる一般的な全文データベースに対して適用可能であることは言うまでもない。

【0085】

【発明の効果】以上説明したように、本発明によれば、アドレス付与手段が、文書データ中の各文字に連続するアドレスを付与し、アドレステーブル作成手段が、前記文書データ中の各文字と後続する文字とで構成される合計k文字（kは2以上）の疑似単語を作成し、各疑似単語の先頭文字列に付与される各アドレスを、対応する疑似単語の文字コード順にアドレステーブルへ記憶する。そして、最下層構築手段が、固有の文字コードを有する疑似単語を固有疑似単語として前記インデックステーブル群中の最下層テーブルへ文字コード順に記憶するとともに、各固有疑似単語に前記アドレステーブル中の各アドレスを対応付ける。さらに、階層化手段が、前記インデックステーブル群中の最上層テーブルに記憶される固有疑似単語数が予め設定された数より大である場合、前記最上層テーブルを略均等に分割するように複数の固有疑似単語を抽出し、前記最上層テーブルの上層のテーブルへ前記複数の固有疑似単語を文字コード順に記憶する。インデックステーブル群は、このような階層構造を有するので、検索時において、十分な検索速度を得ることができるという効果がある。また、インデックステーブル群中の各テーブルに記憶される疑似単語は、固有の文字コードを有する固有疑似単語であるので、インデックステーブル群のサイズが小となるという効果を得ることができる。

【図面の簡単な説明】

【図1】本発明の一実施例による全文データベースシステムの概略構成を示すブロック図である。

【図2】同実施例による全文データベース構築処理の流れを示すフローチャートである。

【図3】同実施例で用いられる文書データの一例を示す図である。

【図4】マーク付加処理が行われた文書データの一例を示す図である。

*

【図3】

多角経営で有名なA社（東京都）は、絶滅の危機にさらされている動物を保護する京都の団体に多額の寄付を行う一方、経常利益の10%に相当する額の裏金を、D銀行経由でE国のペーパーカンパニーへ入金していた。1992年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、A社は第2次オイルショック以来の経営危機に直面した。

...

world wide economic help can help
world economy ...

JA

EA

【図6】

CT		
ワードアドレス	ワード	
1	world	RII
2	wide	
3	economic	RII
4	help	
5	can	RII
6	help	
7	world	
8	economy	

【図17】

1 2 3 4
経営危機

*【図5】各種アドレスが付与された文書データの一例を示す図である。

【図6】欧文対照テーブルCTの概略構成を示す図である。

【図7】ワードアドレステーブルWATおよびワードリストWLの概略構成を示す図である。

【図8】疑似単語テーブルPWTの概略構成を示す図である。

【図9】仮想アドレステーブルVATおよび欧文基本テーブルEBTの概略構成を示す図である。

【図10】和文疑似単語テーブルJPTの概略構成を示す図である。

【図11】文字アドレステーブルCATおよび和文最下層テーブルJ2-1の概略構成を示す図である。

【図12】和文最上層テーブルJ2-2および1文字テーブルJ1等の概略構成を示す図である。

【図13】本発明の一実施例による全文データベースシステムにおける検索処理の流れを示すフローチャートである。

【図14】同システムにおける検索処理の流れを示すフローチャートである。

【図15】検索メニューの一例を示す図である。

【図16】検索結果の一例を示す図である。

【図17】検索文字列の一例を示す図である。

【符号の説明】

3 構築処理装置（アドレス付与手段、アドレステーブル作成手段、最下層構築手段、階層化手段）

6 入力装置（入力手段）

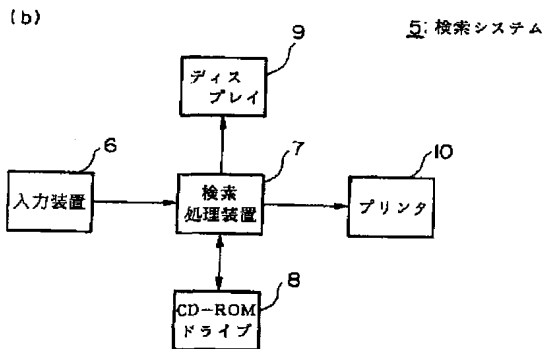
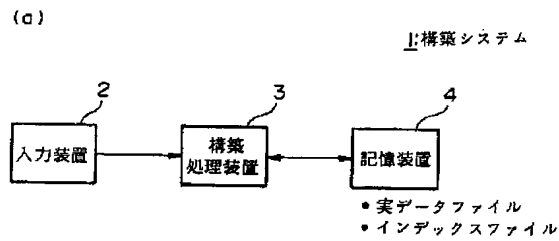
7 検索処理装置（分割手段、検索手段）

CAT 文字アドレステーブル（アドレステーブル）

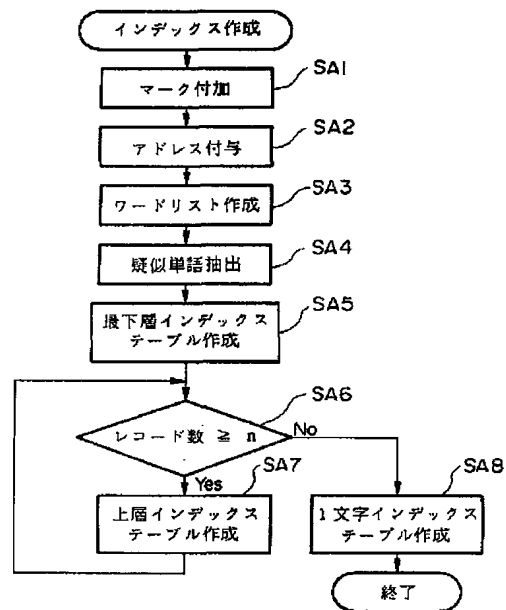
J2-1 和文最下層テーブル（最下層テーブル）

J2-2 和文最上層テーブル（上層テーブル）

【図1】



【図2】



【図4】

多角経営で有名なA社（@東京都@）は、絶滅の危機にさらされている動物を保護する@京都@の団体に多額の寄付を行う一方、経常利益の10%に相当する額の裏金を、D銀行経由でE国のペーパーカンパニーへ入金していた。1992年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、A社は第2次オイルショック以来の経営危機に直面した。

...

world wide economic help can help
world economy ...

JA

EA

【図8】

仮想アドレス	疑似単語	
A-1	ca	R9
A-2	an	R9
B-1	ec	R9
B-2	co	R9
B-3	on	R9

【図5】

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

多角経営で有名なA社（@東京都@）は、絶滅の危機にさらされている動物を保護する@京都@の団体に多額の寄付を行う一方、経常利益の10%に相当する額の裏金を、D銀行経由でE国のペーパーカンパニーへ入金していた。1992年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、A社は第2次オイルショック以来の経営危機に直面した。

...

1(317) 2 3 4 5 6

world wide economic help can help

7 8

world economy ...

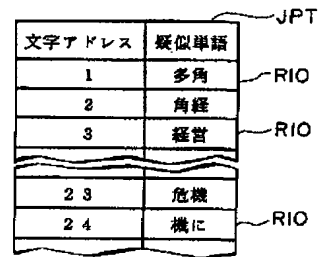
JA

EA

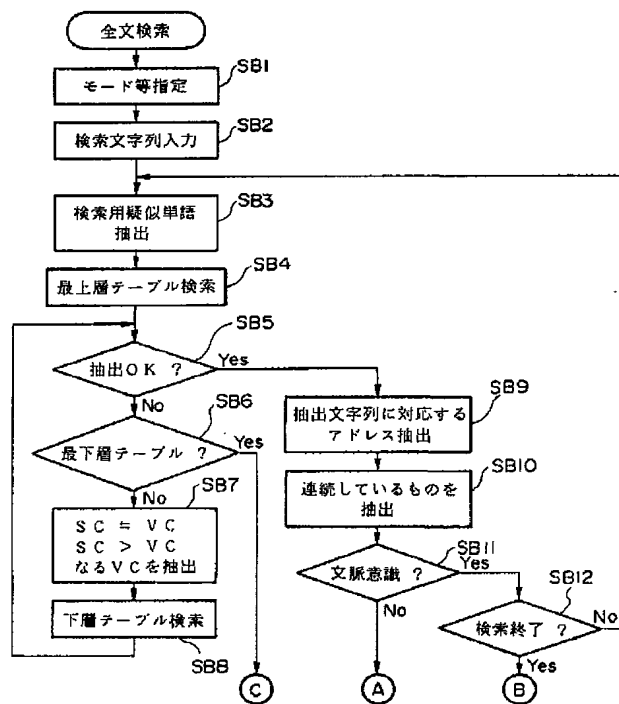
【図9】

	仮想アドレス	ポインタ	サイズ	疑似単語	
	A-2	→	1	an	R7
	A-1	→	1	ca	R7
	B-2	→	2	co	R7
	C-2	→	1	de	R7
	E-3	→	2	ec	R7
	B-1	→			R7
	C-1	→			R7

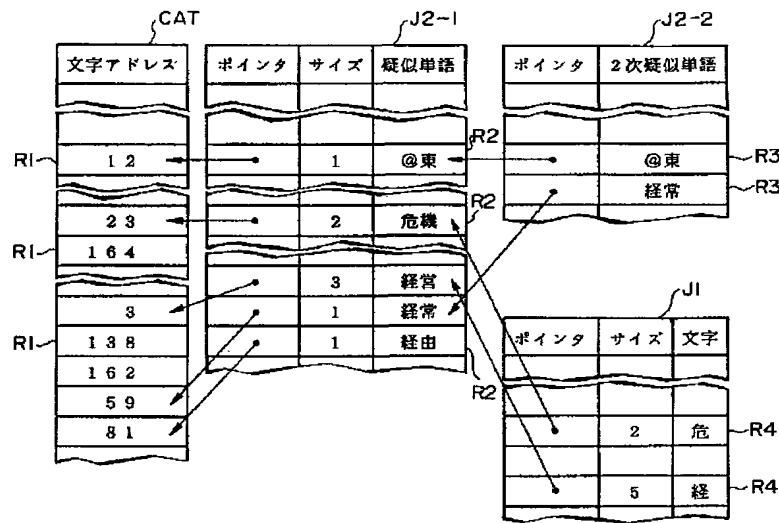
【図 10】



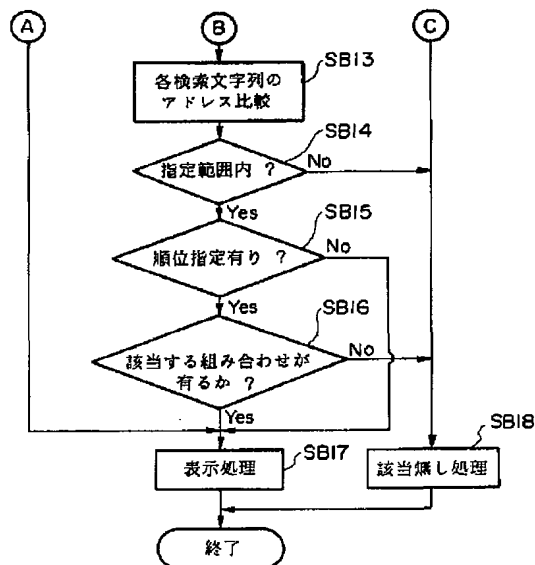
【图 13】



【図12】



【図14】



【図15】

検索モード [2]	順位指定 [1]	距離 [2 0]
1 : 通常検索	1 : 指定なし	
2 : 分脈意識モード	2 : 入力順	
検索文字列:		
[]	[] 件
[]	[] 件
[]	[] 件
分脈意識 [経営, 危機]	[1] 件
	[] 件

【図16】

検索結果詳細
<p>多角経営で有名なA社（東京都）は、絶滅の危機にさらされている動物を保護する京都の団体に多額の寄付を行う一方、経常利益の10%に相当する額の裏金を、D銀行経由でE国のペーパーカンパニーへ入金していた。1992年に、こうしたマネーロンダリングの事実が発覚すると、当時の全経営陣が失脚し、A社は第2次オイルショック以来の経営危機に直面した。...</p>

フロントページの続き

(72)発明者 奈良 雅人
東京都台東区台東一丁目5番1号 凸版印
刷株式会社内

(72)発明者 濱谷 群二
東京都台東区台東一丁目5番1号 凸版印
刷株式会社内